State of California
Office of the State Chief Information Officer

Enterprise Architecture Standard

Geocode Standard
Reference Model Type and ID No: SRM 74.742.590.1
Status: Proposed
Analysis: OCIO Geographic Information Systems and Enterprise Architecture
Effective Date: mm/dd/yyyy
Next Review Date: mm/dd/yyyy
Approved By: Office of the State Chief Information Officer (OCIO)

Introduction

Many State data resources consist of records with a locational component, such as street address, ZIP code or County name. While a sense of place can be inferred from these descriptive components, digital analysis and data display require the assignment of latitude and longitude, or X,Y coordinates, to a point identifying the location of each record. This is accomplished by the process of geocoding, in which each record's locational data is compared against reference data records with known latitude and longitude. The reference record's coordinates are then assigned to the matching data record.

Once the data records are geocoded, they may be compared to other events and information occurring at identical or nearby coordinates. Patterns and relationships can then be appreciated which are not revealed by simple tabular data. These may include:

Issue	Description
Fraud	Potential fraud cases identified when the normal distribution of government funds (e.g. unemployment insurance or public health assistance) is far exceeded at a given address.
Health Patterns	Disease clusters or abnormal distribution patterns visualized from point data representing instances of infection or patient addresses
Environmental Patterns	Evaluation of environmental concerns given the concentration of toxic release reports within a range of addresses
Taxation	Assessment of fair tax collection and levees given the normal distribution of taxation at a set of addresses
Economic Development	Strategic planning of business development and growth given employer locations and access to business tax incentives
Government Services	Tailoring of government services to demographic characteristics identified for a set of addresses

Geocoded data resources have the uniquely identifying element of latitude and longitude, which permits addition of related ancillary information, such as the county or legislative district, to a data record. Moreover, if the input location data are standardized as part of the geocoding process, the final result is a single collection of normalized addresses, street names, cities and ZIP codes. Good geocoding technique can reduce human error and effort, resulting in more effective management of our State data resources.

Geocoding workflow consists of several steps, including data cleaning, standardization, and matching. Standardization is of particular importance, in that it identifies address components and converts them into a correct format to increase geocoding match rate.

Geocoding has three main components: 1) Reference Data, 2) Address (or source) Data, and 3) Software (or geocoding applications):

1. **Reference Data**

Reference data are the underlying geographic base files containing georeferenced features, which the geocoding software uses to match the input (source) data. The base files are most often address point locations based on parcel centroids, street centerline data, ZIP code data or place name locations. Reference data are available from both public and private sources. Public data sources include the US Census Bureau TIGER Line files or the US Postal Service data. In order to obtain the best possible geocodes, reference data should be current, complete and accurate.

2. **Address (or Source) Data**

Address or source data is the descriptive place data which will be geocoded. For State datasets, this is usually a physical street address (e.g., 1325 J Street). Nearly every department in State government has a dataset of street address locations for its own building locations and outlets for public access. Source data can also be incident or event locations (e.g., a fire, or vehicle accident), locations of equipment and facilities (e.g., medical supply stockpile locations, hospitals), and monument locations (e.g., 0.1 miles north from intersection of Hwy 49 and Interstate 80 on Hwy 49).

Source data can contain errors in address formatting or transposed elements (e.g. numbers entered incorrectly or transposed, misspelled addresses, nonexistent addresses and non-standardized address elements (e.g., Boulevard vs. BLVD). These errors should be corrected as part of a data cleansing process prior to the onset of geocoding.

In some instances, address data and resulting geocodes may be part of protected personally identifiable information (reportable disease cases, for example). In this case, such data must be securely administered in full compliance with all applicable federal, state, or local privacy laws and regulations, including but not limited to the Health Insurance Portability and Accountability Act (HIPAA)."

Note: In some instances, address data and resulting geocodes may be part of protected personally identifiable information (e.g., reportable disease cases). It is a State of California practice to recognize that a street address geocode constitutes personal “identifiable” information and therefore must be securely administered in full compliance with all applicable federal, state, or local privacy laws and regulations, including but not limited to the Health Insurance Portability and Accountability Act (HIPAA).

3. **Software (or geocoding application)**

The geocoding software or application consists of several components and performs many functions. The main component is the geocoding algorithm, which identifies features in the reference dataset and uses them to match the input data and assign X and Y. Various geocoding applications use different algorithms to arrive at the final match. The basic function is to parse the source data into defined elements for better understanding and matching. In a physical address example this parsing includes the address number, street prefix, street name, the street suffix, street direction, street type, city name, ZIP code, ZIP+4, and state. The software then performs a probabilistic record linkage with a statistically valid form of fuzzy logic to score how well the source data can be matched to the reference data. This type of matching allows for reviews of “almost” matches, scoring thresholds, index tuning, best match and/or candidate matching.

Geocoding software can also perform a standardization process, whereby the source data is standardized to selected guidelines. This process increases the match rate potential of the source data and provides for a common storage taxonomy as well. Geocoding software can return an X, Y location, the standardized address, a match score, a match sequence (the reference data it matched to), and ancillary data as required.

Standard Requirements

The following standard is approved for State of California geocoding methods. This standard is a “process” standard. The description below identifies the appropriate steps to comply with the state standard for managing address data.

Step 1: Define Input or Source Data – Field Definitions

It is the standard of the State of California for agencies or programs collecting and maintaining location data to store a minimum of the following fields for descriptive location or address information. It is acceptable to modify the field definitions to meet specific business needs (change field names or lengths, split address sub-components into separate fields). However, it is required that each geocoded record have all of the fields below. Departments needing to store more detailed address information may consult the URISA Address Standard (http://www.urisa.org/about/initiatives/addressstandard).

Field	Type	Description
First Address Field	Text	Street number, and name, intersections and place names acceptable. Example: 1325 J Street, 7th & J Street, State Capitol Building
Second Address Field	Text	Building, floor, mail stop, PO Box, suite, etc. Example: Suite 1600
City	Text	City Name Example: Sacramento
State	Text	State Abbreviation Example: CA
ZIP Code	Text	ZIP Code Example: 95814
ZIP 4 Extension (optional)	Text	ZIP Code + “-“ and 4 digit extension if available Example: 95814-1234.

Step 2 – Standardization and Validation Process

Once data is assembled in the proper format, it is the standard of the State of California for geocoding processes to first standardize and validate the accuracy of street address records according to current United States Postal Service (USPS) address data by means of a USPS Coding Accuracy Support System (CASS) Certified software product (http://www.usps.com/ncsc/addressservices/certprograms/cass.htm)(http://www.usps.gov/cass.htm) CASS-certified software can be part of a single geocoding package, a stand alone software product and/or a web service.

Step 3: Composite Geocoding Process

Geocoding software may use one or more reference datasets. A composite geocoding service will rely on multiple reference datasets, and starting with the best quality reference data, attempt to match the source data record. If no match is reached in the best quality reference data set, the service will turn to the next reference data set and attempt to perform the same function. This repeats in a cascading fashion, most to least accurate, until a match is found. A composite geocoding service performs the same function as a single reference geocoding service, and in addition specifies the reference layer to which it matched. Record level analysis can reveal the rate for best quality versus lower quality matches; additionally, a location is identified for each source data record, rather than leaving some unmatched. The State of California geocoding application standard is a composite service to ensure the greatest possible number of record matches.

Step 4 – Keep Post Geocoded Address Fields

It is the standard of the State of California for agencies to keep the following fields from the results of the geocoding process. It is acceptable to modify the field definitions to meet specific business needs (e.g., change field names or lengths, split address sub-components into separate fields). However, each geocoded record must carry all of the fields below. Typical industry geocoding software will return all of these fields.

Field	Type	Width	Description
Geocoding Score	Integer	3	A score rating the quality of address match (geocode) to a reference data set. Business rules for minimum match score are the responsibility of the sponsoring agency for defining. Example: 100
Geocoding reference data	Text	50	A string indicating the source of the reference data to which the address was matched Example: 1-TA_Points_ZIP_0708
Standard Address	Text	50	The standardized and validated address captured for address data quality and management (output of Address 1). Example: 2575 Sand Hill Rd
Standard City Name	Text	30	The standardized and validated city name captured for data quality and management (output of City). Example: Davis
Standard ZIP	Text	5	The standardized and validated ZIP Code captured for data quality and management (output of ZIP) Example: 95827.
Longitude	Floating decimal	8	The x-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need. Example: -120.554987
Latitude	Floating decimal	8	The y-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need. Example: 37.491958

Step 5 – Final Database Management

Agencies are able to maintain these fields in data models suiting their appropriate needs, as long as a minimum number of the above fields are maintained at the record level.

Definitions

Below are definitions pertinent to the geocoding processes that are included in the SIMM 58C, Enterprise Architecture Glossary:

Geocode – A standardized representation for a location given a textual description of the location like an address, ZIP Code, or place name. The standardized representation is typically an X, Y coordinate and/or a latitude and longitude. Generally speaking, geocode refers only to a street address text description of place.

Composite Service – A service used to provide a geocode address whereby multiple layers of address data are used in a hierarchy to achieve maximum accuracy.

Coding Accuracy Support System (CASS) – The CASS is the United States Postal Service Coding Accuracy Support System. The CASS enables the USPS to evaluate the accuracy of address-matching software programs and provide grades for vendors of software. In addition, the vendors then have an ability to change and modify their software to increase the accuracy of address-matching functions. Many currently available software packages meet the CASS standard.

Authorities

Section 11545 of the Government Code (b) The duties of the State Chief Information Officer shall include, but are not limited to all of the following: (2) Establishing and enforcing state information technology strategic plans, policies, and standards, and enterprise architecture.

Implementation

This EA Standard applies to all new data system development for IT projects approved after July 1, 2010, that are initially funded in the Budget Act of 2010.

For systems that are already in place, state agencies should review the EA Standard, and incorporate implementation or retrofit plans into their Agency Information Management Strategy.

Exceptions to this EA Standard may be submitted to the OCIO by following the “OCIO EA Compliance Component Instructions” found in the SIMM 58A, Enterprise Architecture Developers Guide.

Data stored in individual desktop productivity tools, such as spreadsheets, is not subject to this EA standard. However, agencies interested in geocoding such data for mapping purposes are encouraged to follow the EA Standard and associated EA Practice.